

FANTOM: Functional and Taxonomic Analysis of Metagenomes

User Manual

1- FANTOM Introduction:

a. What is FANTOM?

FANTOM is an exploratory and comparative analysis tool for Metagenomic samples.

b. What is not FANTOM?

FANTOM is not a pipeline for high throughput DNA sequence processing.

c. What it does?

FANTOM reads tabular input files including taxonomically and functionally annotated (KEGG Orthology and COG) abundance data and metadata information regarding metagenomic samples obtained from different environments. It clusters the abundance data in more general hierarchy levels with respect to the corresponding database hierarchies (e.g NCBI Taxonomy: species,genus...phylum or KEGG: KO, pathway etc.) and applies statistical operations including multivariate analysis, hypothesis testing and correlation analysis to whole or subsets of the metagenomic abundance data. It finally provides with several result tables and visualization options that allocate FANTOM's position in the tertiary analysis step of a typical Next Generation Sequence processing pipeline.

2- Installation:

FANTOM can run on Windows, Linux or Mac OSX systems.

a. Installer packages:

Win32:

Double click on 'fantom_v101_win32.exe' and follow the instructions for the installation.

Linux:

- Copy and untar the file 'fantom_v101_linux.tar.gz' into the directory where you want to run FANTOM by typing the below commands on terminal:

```
> cp fantom_v101_linux.tar.gz /path/to/fantom/directory'
```

```
> tar -xvzf fantom_v101_linux.tar.gz'
```

- Get into the directory FANTOM/fantom:

```
> cd FANTOM/fantom/
```

- Run FANTOM:

```
> ./fantom
```

Mac OSX:

- Double click on the file 'FANTOM_v101_mac.tar.gz' and start the installer by again double clicking on the uncompressed file 'FANTOM_v101_mac.dmg'. From the installer window; drag and drop FANTOM into the Applications folder.

b. Running from source:

Requirements:

- python 2.7+
- numpy
- scipy
- matplotlib
- wxpython
- Storm

After finding the most stable way to install the above packages for running FANTOM from source; download the codes from 'http://www.sysbio.se/FANTOM/source_codes/'.

Untar the file 'fantom_src.tar.gz' by typing:

```
> tar -xzf fantom_src.tar.gz
```

Change directory into fantom_src and type:

```
> python fantom.py
```

3- Input files:

a. Metadata Input

Name	Country	Gender	Age	BMI	IBDCohort
MH0001	Denmark	female	49	25.55	N
MH0002	Denmark	female	59	27.28	N
V1.CD-15	Spain	female	34	19	CD
NO2.UC-11	Spain	female	34	18.68	UC
O2.UC-12	Spain	male	43	21.6	UC
O2.UC-13	Spain	female	68	23.38	UC

Table 1 – Input metadata file

Input metadata file is a tab separated file involving samples' meta information in each column under appropriate column headers. First column should always be spared for the 'Name' of the corresponding sample. Data types are automatically recognized for each meta property. Ambiguous or absent data should be entered as an empty cell for categorical and 0 (without any other character) for integer or decimal values.

b. Feature Abundance Input

KO	MH0009	MH0008	MH0001	MH0003	MH0002				
K01369	134	125	37	709	954	77	134	12	1473
K01365	0	0	0	0	0	2	3	0	0
K01364	1950		47	754	329	175	3733		525
K01366	0	0	0	78	0	0	0	0	0
K01361	1073		143	173	3944		1015		391

Table 2 - Functional abundance input

Input feature abundance file is a tab separated file having functional or taxonomic database categories ('KO','COG','PFAM','TIGRFAM','TAXON') as column headers and the corresponding database accession codes / identifiers as the first column content and 'absolute' abundance values as the content of the corresponding sample columns.

c. Functional Database Input:

Cellular Processes	Cell Motility	Bacterial chemotaxis	cheC; chemotaxis protein CheC	K03410	
Cellular Processes	Cell Motility	Bacterial chemotaxis	cheY; two-component system, chemotaxis family, response regulator CheY	K03413	
Cellular Processes	Cell Motility	Bacterial chemotaxis	cheB; two-component system, chemotaxis family, response regulator CheB	[EC:3.1.1.61]	K03412
Cellular Processes	Cell Motility	Bacterial chemotaxis	cheV; two-component system, chemotaxis family, response regulator CheV	K03415	
Cellular Processes	Cell Motility	Bacterial chemotaxis	cheZ; chemotaxis protein CheZ	K03414	
Cellular Processes	Cell Motility	Bacterial chemotaxis	rbsB; ribose transport system substrate-binding protein	K10439	
Cellular Processes	Cell Motility	Bacterial chemotaxis	aer; aerotaxis receptor	K03776	
Cellular Processes	Cell Motility	Bacterial chemotaxis	motB; chemotaxis protein MotB	K02557	
Cellular Processes	Cell Motility	Bacterial chemotaxis	mg18; methyl-galactoside transport system substrate-binding protein	K10540	
Cellular Processes	Cell Motility	Bacterial chemotaxis	dppA; dipeptide transport system substrate-binding protein	K12368	
Cellular Processes	Cell Motility	Bacterial chemotaxis	motA; chemotaxis protein MotA	K02556	
Cellular Processes	Cell Motility	Regulation of actin cytoskeleton	ROCK; Rho-associated, coiled-coil containing protein kinase	[EC:2.7.11.1]	K04514
Cellular Processes	Cell Motility	Regulation of actin cytoskeleton	VIL2; villin 2 (ezrin)	K08007	
Cellular Processes	Cell Motility	Regulation of actin cytoskeleton	APC; adenomatosis polyposis coli protein	K02085	
Cellular Processes	Cell Motility	Regulation of actin cytoskeleton	ERK1_2; extracellular signal-regulated kinase 1/2	[EC:2.7.11.24]	K04371
Cellular Processes	Cell Motility	Regulation of actin cytoskeleton	DOCK1; dedicator of cytokinesis 1	K13708	
Cellular Processes	Cell Motility	Regulation of actin cytoskeleton	SSH; slingshot	K05766	
Cellular Processes	Cell Motility	Regulation of actin cytoskeleton	F2R, PAR1; coagulation factor II (thrombin) receptor	K03914	
Cellular Processes	Cell Motility	Regulation of actin cytoskeleton	BDKRB1; bradykinin receptor B1	K03915	
Cellular Processes	Cell Motility	Regulation of actin cytoskeleton	BDKRB2; bradykinin receptor B2	K03916	

Table 3 - Functional database input

Functional database input is a tab separated file involving the hierarchical categories of a functional database where every odd numbered column corresponds to the name and every second column corresponds to the accession number of the functional database. It supports the many to many relationships between the different levels such as in KEGG a Kegg Orthology Group belongs to many pathways and a pathway comprises many Kegg Orthology Groups. If a custom database is given as input for FANTOM, accession codes of the lowest level should match the accession codes given in the functional abundance input file.

4- Graphical User Interface Components:

a. 'Data Import/Configuration' Dialog:

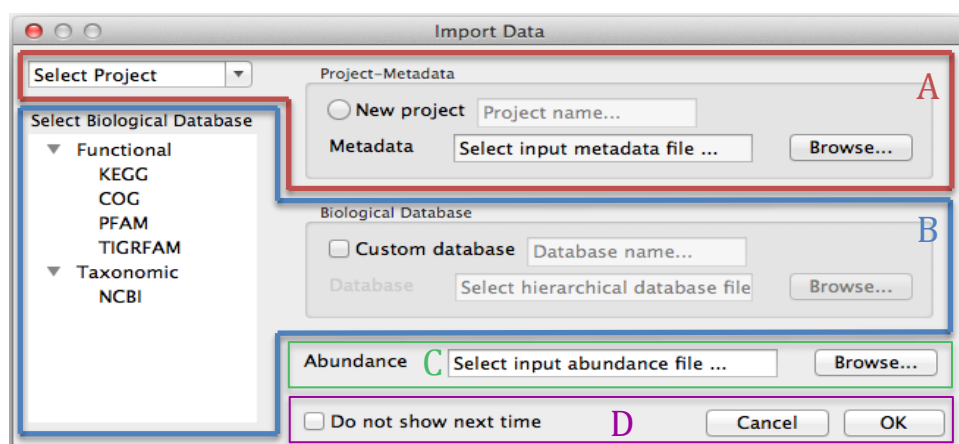


Figure 1 - Data import/Configuration dialog

FANTOM starts with a configuration dialog box where the user enters the initial settings/input files for the metagenomic analysis. User starts the configuration by clicking on the new project radio button and enters the name of the project in the text box. Metadata input file is imported and saved to the local database by clicking on the 'Browse button' in section A of Figure X. Alternatively, if the user have already saved the project, he/she can select a project name by clicking on the arrow near 'Select Project' drop down list. Database type is selected in section B for the biological question of interest. In addition to those built in database options, any

hierarchical database can be imported as input file for FANTOM by clicking on 'Custom database' checkbox, entering the new database name and importing the file by clicking on the 'Browse' button in section B. Feature abundance input file is imported in section C by clicking on the 'Browse' button. A special care has to be taken for the feature abundance file that its column headers have to match with the sample names in the metadata input file. Finally, the user can click on the checkbox in section D not to see this dialog box each time FANTOM is being started.

b. Main Graphical User Interface:

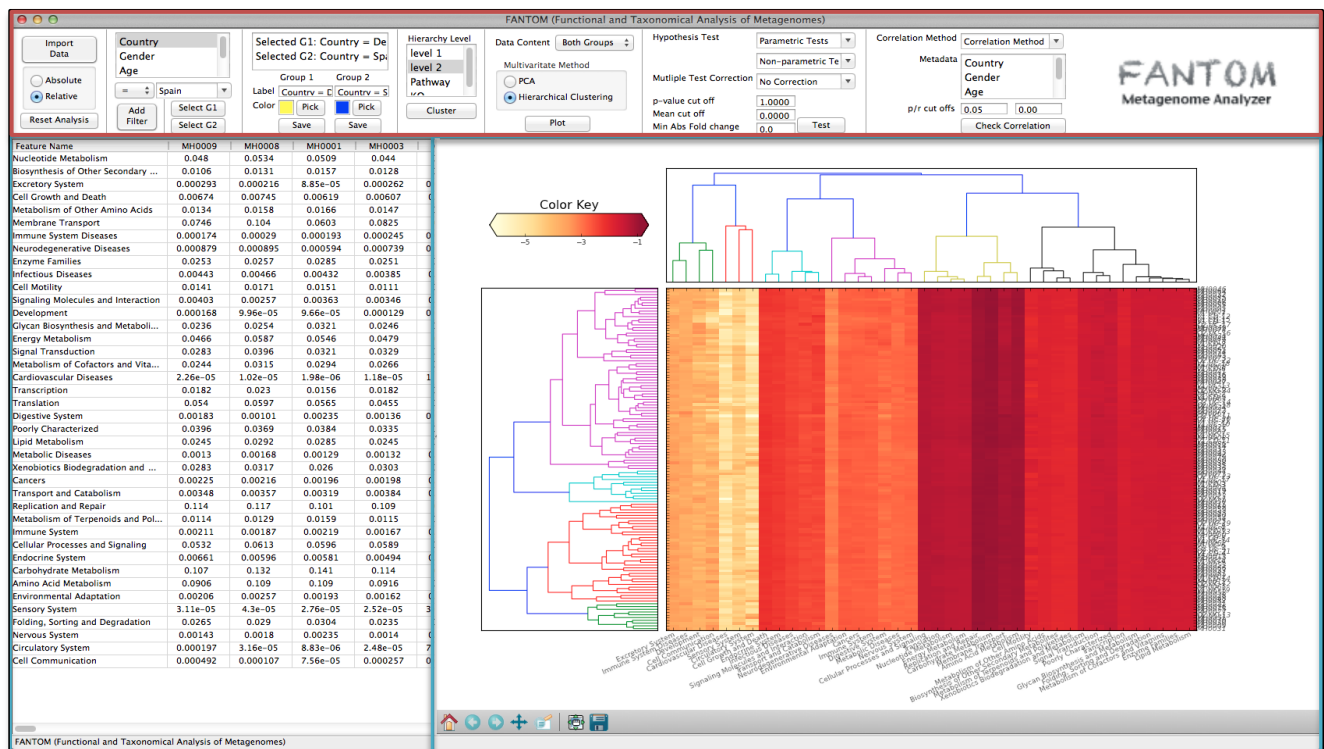


Figure 2 - Main Graphical User Interface

After entering the input options; GUI of FANTOM shows up with its 3 main components: selection and data analysis panel (top), data/results display panel (left), plot panel (right). Functions of each individual component are explained below.

c. Selection and Data Analysis Panel

Data Import:

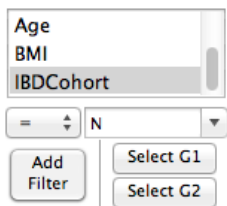
Import Data

☐ Absolute
 ☒ Relative

Show All Data

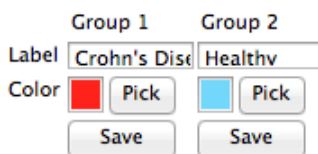
New data sets can be imported to FANTOM during an analysis by clicking on 'Import Data' button. User can then reselect the initial setting from the 'Import Data' dialog box. Metagenomic abundances are typically obtained as counts of matching sequence reads from sequence alignment tools such as blast, HMMER, bowtie etc. and imported to FANTOM as absolute count values. These values can be converted to relative abundances by clicking on the Radio Box above during an analysis and then again can be displayed as absolute counts. Relative abundances are calculated by dividing the count values by the total sum of the counts for each sample. One important issue here is the calculation of relative abundances for higher level database categories where one single functional feature from the lower level category is involved. For example enzymes belonged to a KEGG Orthology group can be involved in many different pathways. In this situation, we simply sum up the relative abundances of the KEGG Orthology level separately for each pathway category it is involved.

Data Selection/Filtering:



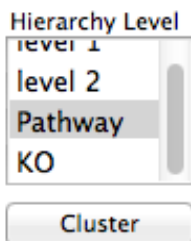
Metagenomic abundance data (either in relative or absolute forms) can be filtered and selected according to their metadata. An appropriate operator for the metadata data type (text or number) can be selected to adjust the filtering. Metadata information for the selected property can be visualized and selected as a filtering/selection option in addition to manually entering the value in the Combo Box in the figure. Current filtering/selection operators are 'contains', '=', '<', '>', '<=', '>='. 'contains' operator allows to perform the filtering/selection by the search term entered in the combobox instead of selecting from the list.

Selected Dataset Display/Labeling:



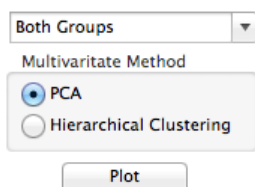
Two datasets to be tested for hypothesis according to their metadata has been selected above and these datasets can be relabeled within this box as well as selecting custom colours for each group to be represented by various plotting options.

Clustering Feature Abundances by Database Hierarchy:



Abundance counts can be summed up by clustering the features in the same hierarchical category group together according to their level in the corresponding database. Selecting the 'Pathway' option in the example on the left will sum up the abundances of KO level counts belonging to individual pathways and list them on the data/results display panel.

Multivariate Analysis/ Data Exploration:



Multivariate data analysis including Principal Component Analysis (PCA) and Hierarchical Clustering can be applied for all data, the two selected groups or for each individual group separately. Corresponding plots are displayed in the plot panel as below by clicking on the 'Plot' button.

Statistical Hypothesis Testing:

Hypothesis Test	Parametric Tests	
	Mann-Whitney Test	
Multiple Test Correction	False Discovery Rat	
p-value cut off	0.2	
Mean cut off	0.00001	
Min Abs Fold change	2	Test

Two previously selected groups can be subjected to Statistical Hypothesis Testing including the parametric tests "Levene's Test", "Bartlett's Test", "Welch's t-test", "Student's t-test" and non parametric "Mann-Whitney U Test". P-values obtained by the tests can be adjusted for multiple testing correction by applying Bonferonni correction or Benjamini-Hochberg False Discovery Rate. Results can be filtered by entering cut-off values for minimum p-value, minimum abundance mean or minimum absolute fold change in means of the two groups. The results are displayed in the data panel as below as well as the plots

Correlation Analysis:

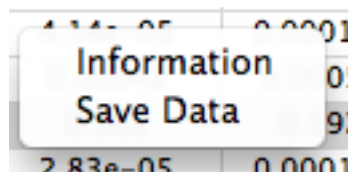
Correlation Method	Correlation Method
Metadata	Country Gender Age
p-value cut off	1.0000
Check Correlation	

Finally correlation analysis can be applied for the selected datasets with a selected metadata property from the above listbox. Again a p-value cut-off can be entered for filtering the results. Correlation analysis results are listed on the data/results display panel. A mouse click on a feature from the results table will plot the scatter plot of the selected metadata against the feature abundances in addition to the regression line fitting the data.

d. Data/Results Display Panel:

	Feature Name	V1.CD-15	V1.CD-12	V1.CD-6	
1	Nucleotide Metabolism	0.189	0.176	0.194	
2	Biosynthesis of Other Secondary Me...	0.083	0.0811	0.0798	
3	Excretory System	0.00312	0.00342	0.00245	
4	Cell Growth and Death	0.0367	0.0413	0.0501	
5	Metabolism of Other Amino Acids	0.083	0.0812	0.0952	
6	Membrane Transport	0.148	0.124	0.184	
7	Immune System Diseases	0.00149	0.00169	0.00179	
8	Neurodegenerative Diseases	0.00445	0.00394	0.00594	
9	Immune System	0.000506	0.000567	0.000579	
10	Infectious Diseases	0.0128	0.0138	0.0167	
11	Cell Motility	0.0237	0.0384	0.0575	
12	Signaling Molecules and Interaction	3.73e-05	0.000125	0.000327	
13	Development	0.000378	0.000495	0.000654	
14	Glycan Biosynthesis and Metabolism	0.145	0.173	0.157	
15	Cell Communication	2.55e-05	0.000145	0.000391	
16	Signal Transduction	0.00412	0.00462	0.00482	
17	Metabolism of Cofactors and Vitami...	0.189	0.168	0.183	
18	Cardiovascular Diseases	7.91e-05	6.36e-05	3.85e-05	
19	Transcription	0.0592	0.0692	0.0689	
20	Translation	0.142	0.121	0.155	
21	Digestive System	0.0131	0.0144	0.0124	
22	Lipid Metabolism	0.0673	0.0732	0.0832	
23	Metabolic Diseases	0.0144	0.0133	0.0142	
24	Xenobiotics Biodegradation and Me...	0.0494	0.0437	0.0511	
25	Cancers	0.00471	0.00677	0.00827	
26	Transport and Catabolism	0.0463	0.066	0.0395	
27	Replication and Repair	0.144	0.161	0.175	
28	Metabolism of Terpenoids and Poly...	0.0261	0.0226	0.0285	
29	Endocrine System	0.0165	0.0177	0.0202	
30	Amino Acid Metabolism	0.227	0.21	0.221	
31	Environmental Adaptation	0.00506	0.00335	0.00652	
32	Sensory System	0.000429	0.000948	0.00105	
33	Folding, Sorting and Degradation	0.113	0.101	0.119	
34	Carbohydrate Metabolism	0.365	0.387	0.37	
35	Energy Metabolism	0.28	0.241	0.269	

Table 4 - Data/Result Display Panel



By right clicking on each functional feature or taxon name the user can either:

- 1- Display information such as the distribution of abundance counts as a histogram, Shapiro Wilk's normality test result or the lineage of the feature in the hierarchical database OR;
- 2- Save the results into a text file.

e. Plot Panel

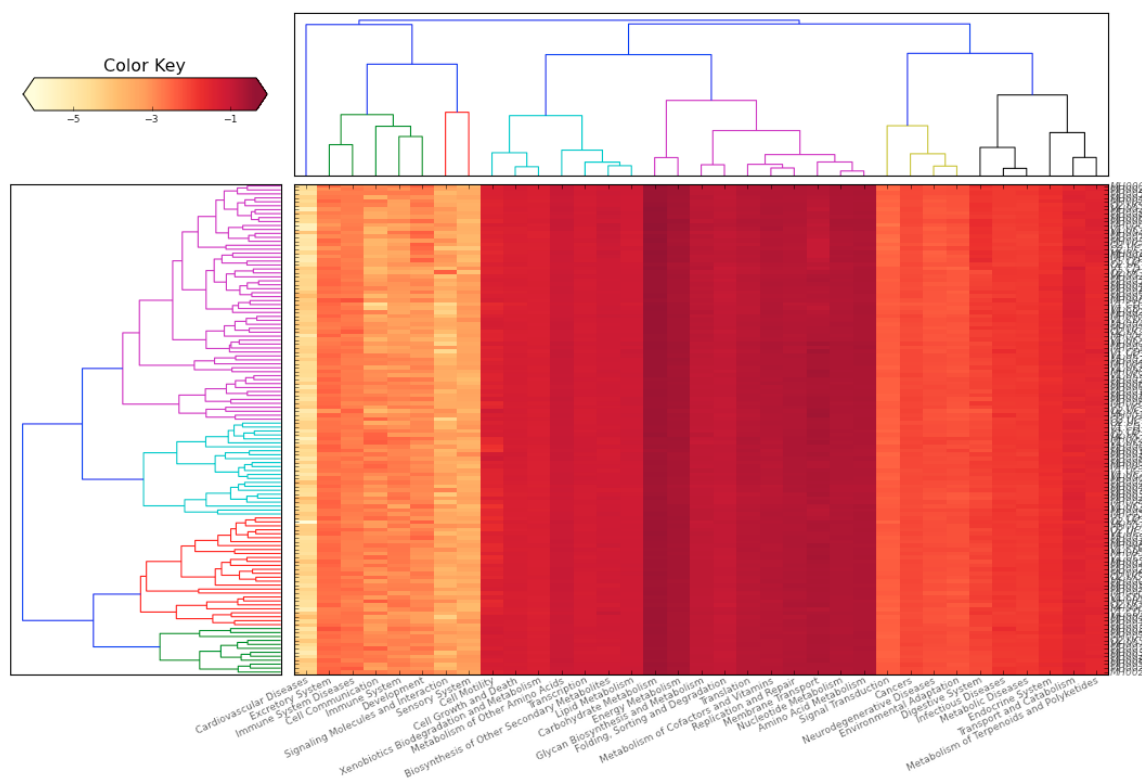


Figure 3 – Plot panel



Right clicking on a 'Bar chart' pops up a menu displaying other plotting options. One of the 'Pie','Box','Area' or 'Bar' chart options can be selected and plotted with the previously assigned labels and colors for both groups.



Figure 4 – Matplotlib interactive navigation panel

Toolbox at the bottom of plot panel (Figure 2) is the default matplotlib interactive navigation panel where user can reset the changes (1), undo (2) or redo (3) the changes, move (4), zoom in the plot (5) and can finally save it on the disk (7). Subplot configuration tool (6) of the interactive navigation toolbox will not work in FANTOM since plots are fixed on the plot panel.

f - All Outputs:

Result Tables:

	Feature Name	Test S	p-value	Mean 1	Error 1	Mean 2	Error 2	Abs Fold Ch
1	Biosynthesis of siderophore group nonri...	10	0.0943	0.00128	0.000193	0.00337	0.00102	2.63
2	Synthesis and degradation of ketone bod...	4	0.0585	0.000538	0.000146	0.00142	0.000158	2.65
3	C5-Branched dibasic acid metabolism	11	0.112	5.35e-05	1.52e-05	0.00027	6.57e-05	5.05
4	Sphingolipid metabolism	2	0.0546	0.00419	0.000581	0.00172	0.000184	2.44
5	Bile secretion	5	0.0537	3.29e-05	5.65e-06	0.000109	1.51e-05	3.32
6	p53 signaling pathway	1	0.0592	1.57e-05	4.23e-06	0.000385	0.000126	24.6
7	beta-Lactam resistance	9	0.0843	0.00108	0.000214	0.00255	0.000312	2.37
8	Inositol phosphate metabolism	6	0.059	0.000505	7.88e-05	0.00104	0.000124	2.06
9	RIG-I-like receptor signaling pathway	8	0.0703	1.49e-05	4.6e-06	8.73e-05	2.76e-05	5.86
10	Flavonoid biosynthesis	4	0.0563	7.13e-05	5.61e-05	0.000623	0.000139	8.74
11	Arrhythmogenic right ventricular cardiom...	5	0.0523	4.91e-05	5.92e-06	2.3e-05	3.44e-06	2.14
12	Endocytosis	11	0.111	6.39e-05	2.15e-05	0.000135	2.57e-05	2.11
13	Non-homologous end-joining	15	0.172	6.62e-06	1.87e-06	1.79e-05	3.32e-06	2.7
14	Wnt signaling pathway	15	0.177	4.92e-06	3.04e-06	1.69e-05	5.1e-06	3.44
15	ECM-receptor interaction	15	0.176	7.78e-05	2.92e-05	0.000161	3.39e-05	2.07
16	Axon guidance	4	0.0475	0.00029	5.51e-05	0.000898	0.000158	3.09
17	Phagosome	1	0.0739	2.91e-06	1.07e-06	3.91e-05	7.23e-06	13.5
18	Complement and coagulation cascades	7	0.0648	0.000364	1.25e-05	0.000984	0.000141	2.71
19	Focal adhesion	3	0.051	0.000115	5.76e-05	0.0012	0.000249	10.4
20	Fatty acid metabolism	6	0.0577	0.00193	0.000348	0.00471	0.00057	2.44
21	Dorso-ventral axis formation	0	0.141	6.56e-05	7.01e-06	1.44e-05	2.53e-06	4.55
22	Small cell lung cancer	12.5	0.126	2.3e-06	1.99e-06	1.2e-05	7.22e-06	5.22
23	Carbohydrate digestion and absorption	0	0.106	0.000573	0.000176	0.00364	0.000668	6.35
24	Carotenoid biosynthesis	10	0.1	4.37e-05	3.74e-05	0.000117	2.6e-05	2.67
25	Prion diseases	7	0.0594	8.97e-05	2.23e-05	3.98e-05	9.54e-06	2.25
26	Styrene degradation	4	0.0447	0.000265	4.45e-05	0.000554	5.11e-05	2.09
27	Betalain biosynthesis	13	0.136	1.62e-05	5.87e-06	6.37e-06	2.85e-06	2.54
28	Indole alkaloid biosynthesis	10	0.0992	2.86e-06	2.07e-06	1.53e-05	5.08e-06	5.34
29	Stilbenoid, diarylheptanoid and gingerol ...	6	0.0512	1.93e-05	5.31e-06	6.26e-05	1.04e-05	3.24
30	Hepatitis C	6	0.0629	9.88e-06	6.61e-06	8.31e-05	2.48e-05	8.41
31	Pathogenic Escherichia coli infection	2	0.0482	0.000592	9.62e-05	0.00022	2.97e-05	2.69
32	Aminobenzoate degradation	10	0.0979	0.000337	9.12e-05	0.000708	9.14e-05	2.1
33	Renal cell carcinoma	3	0.0534	0.000143	3.15e-05	0.000389	4.69e-05	2.71
34	Taste transduction	7	0.0566	0.000634	9.17e-05	0.000308	3.66e-05	2.06
35	Parkinson's disease	9.5	0.0912	5.22e-06	3.18e-06	5.8e-05	2.05e-05	11.1
36	Tight junction	14	0.157	2.5e-06	5.54e-07	2.02e-05	9.48e-06	8.08
37	Neuroactive ligand-receptor interaction	12	0.119	3.09e-05	1.74e-05	0.000165	5.34e-05	5.35

Table 5 - Statistical Hypothesis Testing result list

	Feature Name	Corr. coefficient	p-value
1	N-Glycan biosynthesis	-0.281	0.00158
2	Vibrio cholerae infection	-0.207	0.0213
3	Thiamine metabolism	0.181	0.0441
4	Benzoate degradation	-0.249	0.00532
5	Ascorbate and aldarate metabolism	0.186	0.0387
6	ECM-receptor interaction	-0.202	0.0246
7	Vitamin B6 metabolism	0.181	0.0441
8	Phagosome	-0.26	0.00351
9	Complement and coagulation cascades	-0.188	0.0367
10	Basal transcription factors	-0.184	0.0405
11	Fatty acid metabolism	-0.206	0.022
12	Adipocytokine signaling pathway	-0.237	0.00803
13	mTOR signaling pathway	-0.245	0.00602
14	Flagellar assembly	-0.182	0.0432
15	Pancreatic cancer	0.184	0.0412
16	Pentose and glucuronate interconversions	0.202	0.0244
17	Chronic myeloid leukemia	0.214	0.017
18	Biosynthesis of siderophore group nonribosomal peptides	-0.191	0.0334
19	Zeatin biosynthesis	0.218	0.0148
20	Naphthalene degradation	0.243	0.00645
21	Aldosterone-regulated sodium reabsorption	-0.27	0.00245
22	Oocyte meiosis	-0.187	0.0371
23	mRNA surveillance pathway	-0.182	0.0436
24	DNA replication	0.195	0.0302
25	Tropane, piperidine and pyridine alkaloid biosynthesis	-0.198	0.0276
26	Betalain biosynthesis	0.231	0.00985
27	Photosynthesis – antenna proteins	-0.18	0.0459
28	Taurine and hypotaurine metabolism	0.212	0.018
29	Biotin metabolism	0.198	0.0272
30	Tyrosine metabolism	-0.187	0.0373
31	Retinol metabolism	-0.204	0.0234
32	Protein digestion and absorption	0.232	0.00959

Table 6 – Correlation analysis results

Multivariate Analysis Plots:

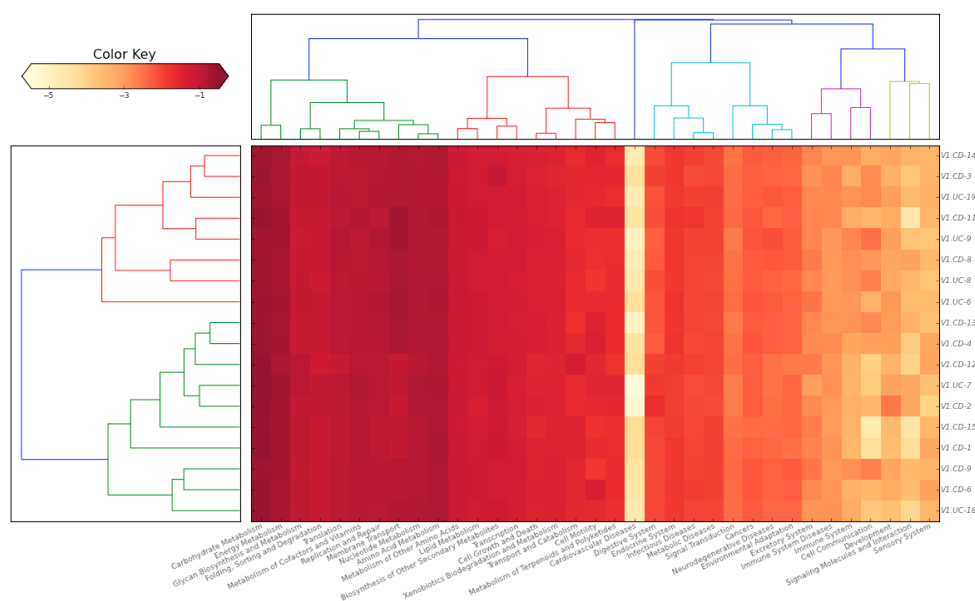


Figure 5 – Hierarchical clustering plot (Corresponding dendograms and the heatmap)

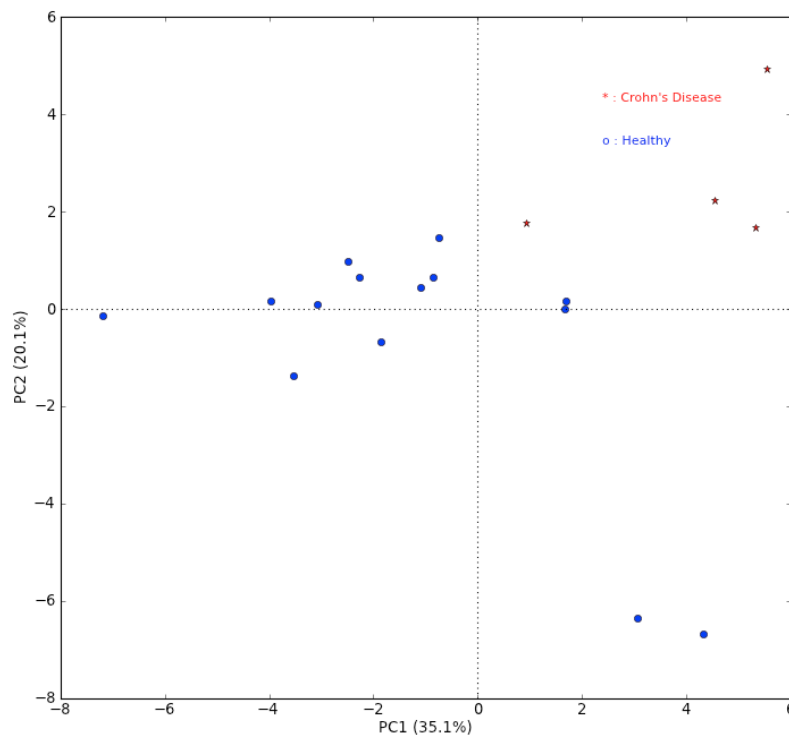


Figure 6 – PCA Plot

Data Subgroup Comparison Plots:

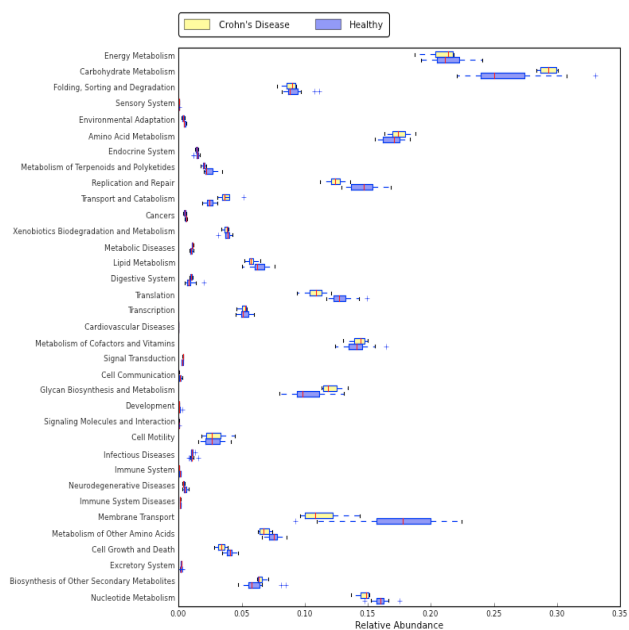


Figure 7 – Box plot

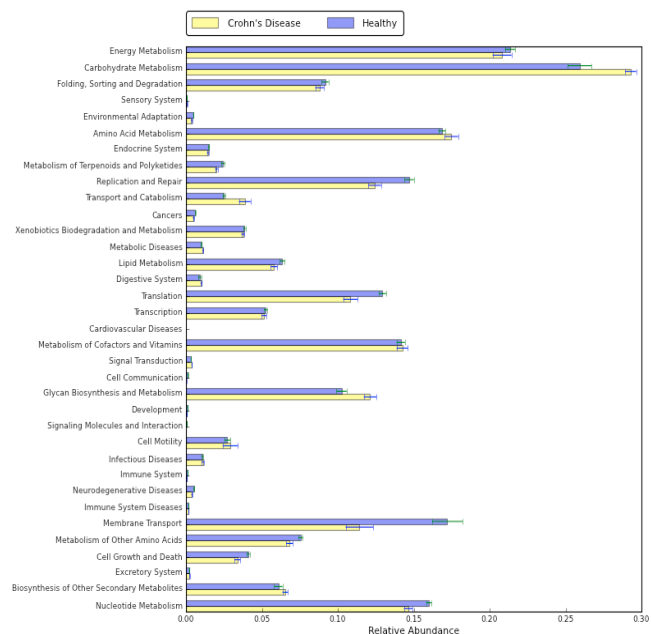


Figure 8 – Bar chart

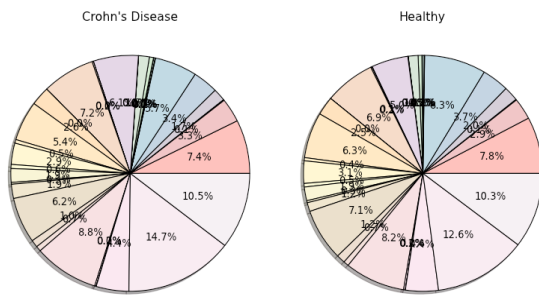


Figure 9 - Pie charts

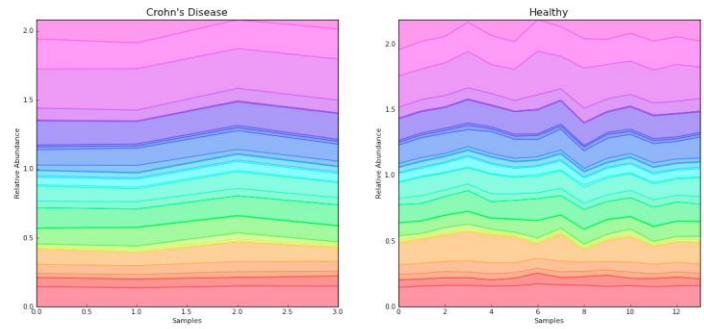


Figure 10 - Area plots

Correlation Analysis Plot:

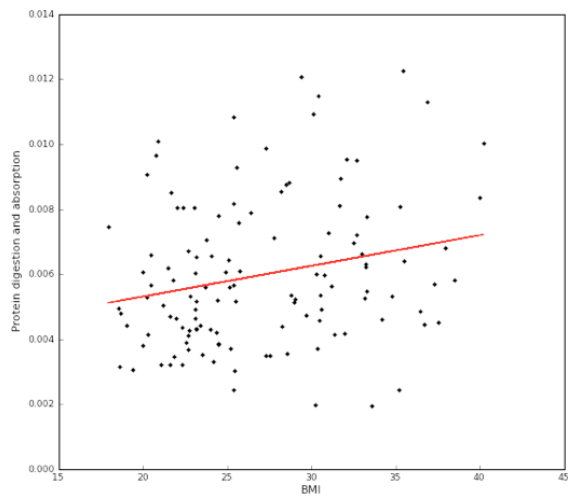


Figure 11 - Metadata - abundance scatter plot with the regression line